

Ching-Tsung (Deron) Tsai

New York, NY | 201-850-3521 | ct2840@nyu.edu | linkedin.com/in/tctsung | github.com/tctsung

WORK EXPERIENCE

Pfizer (formerly Seagen), Seattle, WA (Remote)

Oct. 2023-Present

Senior Software Developer, Oct. 2024-Present

Software Developer, Oct. 2023- Oct. 2024

- **Standardized hierarchical data model transformation** & quality check by developing an R package, actively used by 11 analysts
- Automated extraction of biomarker data and data-migration-related documents from 10+ sources using Python pandas, saving a 4-person development team **32 hours weekly** in building ETL pipelines
- Maintained **8** clinical DBs with >32,000 patient records via Apache Airflow and Docker on Azure. Enabled efficient downstream analysis

Mamon11, New York, NY (Remote)

Jul. 2023-Sep. 2023

Data Science Intern

- Processed online transaction & user behavior data with PostgreSQL & Python polars, successfully streamlining ETL pipeline for **2** clients
- Boosted time-series XGBoost model performance via feature engineering, achieved **13.5% rise in clickthrough rate** & increase revenue

Regeneron, Tarrytown, NY

Jun. 2022-Aug. 2022

Data Engineer Intern

- Transformed ~500G in-house genomic sequences into structured database with regex & NLP, enabling automatic feature engineering
- Built drug target identification pipeline on AWS DNAnexus, discovering **4** sequence patterns using multinomial regression

Chang Gung Memorial Hospital, Inst. of Stem Cell and Translational Cancer Research, New Taipei, Taiwan

Aug. 2019-Feb. 2021

Junior Statistician

- **Designed evaluation metrics** for survival and Cox hazards model, discovering 3 factors to help identify appropriate surgical candidates

PROJECTS

LyricChat: Agentic RAG Chatbot for Emotion-Driven Song Recommendations ([GitHub link](#))

Aug. 2024-Present

- Deployed empathetic chatbot with ReACT framework to recommend mood-matching songs, driving fan engagement for an online singer and placing in **top 10** songs of the week on StreetVoice streaming platform
- Enhanced output relevancy and faithfulness by **22%** with re-ranking & metadata filtering for 3,885-song vector database queries
- Streamlined app deployment by containerizing Qdrant, Ollama & Streamlit UI with Docker, enabling one-click deployment in 3 min

Imbalanced Classification Using Neyman-Pearson Paradigm and Cost-Sensitive Learning ([Package link](#))

Oct. 2021-Apr. 2023

- Fine-tuned Random Forest, XGBoost, LightGBM & CatBoost models for Alzheimer's disease prediction, achieved **26%** class-specific error rate reduction by implementing asymmetric error control for the highest performing CatBoost model (AUC=**0.88**)
- Increased feasibility of multi-class NP algorithm to **15.83x** by integrating the algorithm using R Caret machine learning framework
- Automated **91%** model resampling, visualization, and evaluation process using R ggplot, tidyverse, and published software on CRAN

Feature Tokenized Transformer Classifier with Metagenomic Profiles ([GitHub link](#))

Jan. 2023-Apr. 2023

- Developed Transformer model with PyTorch on HPC system to predict disease risk using microbiome profiles, achieving 0.85 F1 score
- Improved 7% balanced accuracy through regularization & customized deep learning layers using Ray Tune

EDUCATION

New York University, New York, NY

May 2023

M.S. in Biostatistics | Cumulative GPA: 3.9/4.0

- Courses: Deep Learning, Machine Learning, Missing Data, Linear/ Categorical Regression, Survival Analysis, Bayesian/Statistical Inference, Python, R, SQL, SAS

National Chiao Tung University, Hsinchu, Taiwan

Jan. 2019

B.S. in Biological Science and Technology | Major GPA: 3.8/4.0

- Courses: Biostatistics, Machine Learning in Computational Biology, Data Analysis & Statistical Software, Oncology, Python

SKILLS

Programming: Python (polars, pandas, numpy, scipy, seaborn, matplotlib), R (tidyverse, dplyr, ggplot2, MICE), Java, bash, SAS, Excel

Tools: Spark, High-performance computing, Git, SQL, AWS, Airflow, Docker, Singularity, Tableau, Jupyter, Streamlit, Qdrant, Weaviate

Certificate: AWS Certified Cloud Practitioner, Generative AI with LLM (DL.ai), Docker-Hands On-DevOps (Udemy), Python (stanCode)

ML/DL: PyTorch, Hugging face, LangChain, LangGraph, Ollama, Onnx, Instructor, Sklearn, SHAP, Ray tune, TensorFlow, Caret, Glmnet

PUBLICATIONS

[1] Y. Tian, **C. T. Tsai**, Y. Feng (2023) "npcs: Neyman-Pearson Classification via Cost-Sensitive Learning", *CRAN* ([link](#))

[2] R. Garcia, **C. T. Tsai**, G. Ghosh, N. Jandhyala, Y. Feng, H. Rao, C. Thomas, B. Cohen, M. Karajannis, M. Snuderl, J. Allen, D. Segal (2024), "A single-center retrospective study of midbrain gliomas: clinically and radiographically heterogeneous tumors", *Neuro-Oncology Journal* (under review)